bio-protocol Preprint

Not Peer Reviewed
Ask a question online

Phylogenetic distance analysis and clustering

RAS Ricardo A. Segovia

Updated date: Jun 29, 2022

An abbreviated version of this protocol was published in Science Advances in May 2020

Freezing and water availability structure the evolutionary diversity of trees across the Americas

DOI: 10.1126/sciadv.aaz5373

Detailed protocol

```
# chunk phylosor distances matrix and kmeans clustering
# Ricardo Segovia, Institute of Ecology and Biodiversity (ieb-chile.cl)/ 29/06/2022


##CREATING THE COMMUNITY MATRIX
#let's create a dummy variable to have something to operate on
genusXarea <-genusXarea_america
genusXarea$dummy <- 1


#create the genus by site matrix
genus_commat <- tapply(genusXarea$dummy,INDEX=list(site=genusXarea$V1,genus=genusXarea$V2),FUN=sum)
dim(genus_commat)


#fill in the zeros for genera absent from sites
genus_commat[which(is.na(genus_commat))] <- 0
#check and make sure it matches with rows in genusXarea data
sum(genus_commat); dim(genusXarea)



#MAtching phylo and matrix
#let's figure out which genera have multiple accessions and which are in our genus matrix, to figure out which ones we have to deal with
tree_original <- read.tree("R3019.tre") # genus-level phylogeny / genera in more than one continent are labeled genusX-SA (if in South America )genusX-AF (if in Africa)
tree_names_table <- matrix(NA,Ntip(tree_original),4)


for (i in 1:nrow(tree_names_table)){
tree_names_table[i,1] <- unlist(strsplit(tree_original$tip.label[i],split="_"))[1]
tree_names_table[i,2] <- unlist(strsplit(tree_original$tip.label[i],split="_"))[2]
tree_names_table[i,3] <- unlist(strsplit(tree_original$tip.label[i],split="_"))[3]
tree_names_table[i,4] <- unlist(strsplit(tree_original$tip.label[i],split="_"))[4]
}



colnames(tree_names_table) <- c("Order","Family","Genus","Region")
tree_names_table <- as.data.frame(tree_names_table)
rownames(tree_names_table) <- tree_original$tip.label
summary(tree_names_table)
#so, we have lots of genera with multiple accessions, and often found in different regions
#let's focus on those genera that are in the South America table
tree_names_table_sub1 <- tree_names_table[which(tree_names_table$Genus%in%colnames(genus_commat)),]
#let's get rid of empty levels to our genus factor column
tree_names_table_sub1$Genus <- as.factor(as.character(tree_names_table_sub1$Genus))
```

```
summary(tree_names_table_sub1)
#good, so we have removed all genera that are not in phylogeny, but still have a lot of repeated names to deal with
#let's check out those repeated names real quick
tmp <- summary(tree_names_table_sub1$Genus,maxsum=2000)
repeated_genera <- names(tmp)[which(tmp>1)]
#just printing information about the repeated names
for (i in 1:length(repeated_genera)){
print(tree_names_table_sub1[which(tree_names_table_sub1$Genus==repeated_genera[i]),])
}
#it looks like all repeated taxa have one sequence with a _SA appendix, so we can just keep that
#it also looks like the one that is with the _SA appendix is more accurately placed than the one without an appendix (e.g. Capparis, Dacryodes, see names above from Ricardo)
#first get the ones that are not repeated, because we clearly want to keep them
names2keep <- rownames(tree_names_table_sub1)[which(!tree_names_table_sub1$Genus%in%repeated_genera)]
#then get the repeated names
tmp <- rownames(tree_names_table_sub1)[which(tree_names_table_sub1$Genus%in%repeated_genera)]
#only keep ones with south america appendix
tmp2 <- tmp[grep("_SA",tmp)]
names2keep <- c(names2keep,tmp2)
names2exclude <- tree_original$tip.label[which(!tree_original$tip.label%in%names2keep)]
tree_SA <- drop.tip(tree_original,names2exclude)
Ntip(tree_SA)



##change names
newnames <- matrix(NA,length(tree_SA$tip.label))
newnames <- tree_SA$tip.label
newnames <- data.frame(newnames)
newnames ['genus'] <- sapply (strsplit(as.character(newnames$newnames),'_'),"[",3)
tree_SA$tip.label <- as.character(newnames$genus)
write.tree(tree_SA, "sa_tree.tre")


##create America Commat
SA_genus_commat <- genus_commat[,which (colnames(genus_commat) %in% tree_SA$tip.label)]
dim(SA_genus_commat)




##Cluster Analyses


#But, before that, let's go ahead and try to get the full phylosor distance object
phylosor_all <- phylosor.query(tree_SA,SA_genus_commat)
# Warning: one of the input matrices has fewer columns than the number of species in the tree.
#not sure if that could have messed things up...


rownames(phylosor_all) <- colnames(phylosor_all) <- rownames(SA_genus_commat)
phylosor_all_dist <- 1 - phylosor_all
phylosor_all_dist <- as.dist(phylosor_all_dist)


#let's do a simple cluster of this
phylosor_all_cluster <- hclust(phylosor_all_dist,method="average")
phylosor_all_cluster_phylo <- as.phylo(phylosor_all_cluster)
write.tree(phylosor_all_cluster_phylo,"hclustaverage.tre")


####Elbow Analysis. (to select the bes K)
wss <- (nrow(phylosor_all)-1)*sum(apply(phylosor_all,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(phylosor_all,
centers=i)$withinss)
plot(1:15, wss, type="b", xlab="Number of Clusters",
ylab="Within groups sum of squares")


##Try another bayesian approach for select the best K
library(mclust)
```

```
d_clust <- mclustBIC(phylosor_all, G=1:15,
modelNames = mclust.options("emModelNames"))
d_clust$BIC
plot(d_clust)
###Too slow... advanced 4% in a whole night.


# K means and silouethe approach validation



##K-means clustering


set.seed(123)
km.res2 <- kmeans(phylosor_all_dist, 2, nstart = 25)
# k-means group number of each observation
K2<-km.res2$cluster


km.res3 <- kmeans(phylosor_all_dist, 3, nstart = 25)
# k-means group number of each observation
K3<-km.res3$cluster


km.res4 <- kmeans(phylosor_all_dist, 4, nstart = 25)
# k-means group number of each observation
K4<-km.res4$cluster


km.res5 <- kmeans(phylosor_all_dist, 5, nstart = 25)
# k-means group number of each observation
K5<-km.res5$cluster


km.res6 <- kmeans(phylosor_all_dist, 6, nstart = 25)
# k-means group number of each observation
K6<-km.res6$cluster


#saved for K2 to K6
areas_new2<-cbind(areas_new,K2,K3,K4,K5,K6)


##Try plotting ths average silhouette sil_width


silk2<- silhouette(areas_new2$K2,dist(phylosor_all))
si.sumk2 <- summary(silk2)
save(si.sumk2,file="sil_sumk2")
silk3<- silhouette(areas_new2$K3,dist(phylosor_all))
si.sumk3 <- summary(silk3)
save(si.sumk3,file="sil_sumk3")
silk4<- silhouette(clusters7to15$K4,dist(phylosor_all))
si.sumk4 <- summary(silk4)
save(si.sumk3,file="sil_sumk3")
silk5<- silhouette(clusters7to15$K5,dist(phylosor_all))
si.sumk5 <- summary(silk5)
save(si.sumk5,file="sil_sumk5")
silk6<- silhouette(clusters7to15$K6,dist(phylosor_all))
si.sumk6 <- summary(silk6)
save(si.sumk6,file="sil_sumk6")


average_width<-as.vector(c(si.sum2$avg.width,si.sum3$avg.width,si.sum4$avg.width, si.sumk5$avg.width,si.sumk6$avg.width))
rownames(average_width) <- c("K2","K3","K4","K5","K6")


sil_cluster_validation<- as.data.frame(cbind(names,average_width))
class(sil_cluster_validation$names)
```

```
sil_cluster_validation$average_width<-as.vector(sil_cluster_validation$average_width)
sil_cluster_validation$average_width<-as.character(sil_cluster_validation$average_width)


summary(sil_cluster_validation$average_width)
x1 = factor(sil_cluster_validation$names, levels=c("K2","K3","K4","K5","K6","K7","K8","K9","K10","K11","K12","K13","K14","K15"))


pdf("sil_cluster_val.pdf", width = 21, height = 12)


par(mar = c(8,8,4,4))
plot(x1,sil_cluster_validation$average_width,pch=1,type = "b",xaxt="n",
xlab = "Number of Clusters",
ylab="Average Width",mgp=c(4,1,0),
cex.lab=1,cex.axis=1)
title("B)", cex.main=1.5, adj=0,line=0.7)
axis(1,at = seq(1, 14, by = 1), labels = c("K2","K3","K4","K5","K6"))
dev.off()
```

How to cite:(Readers should cite both the Bio-protocol preprint and the original research article where this protocol was used)

1. Segovia, R. A.(2022). Phylogenetic distance analysis and clustering. Bio-protocol Preprint. bio-protocol.org/prep1758.
2. Segovia, R. A., Pennington, R. T., Baker, T. R., Souza, F. C. D., Neves, D. M., Davis, C. C., Armesto, J. J., Olivera-Filho, A. T. and Dexter, K. G. (2020). Freezing and water availability structure the evolutionary diversity of trees across the Americas . Science Advances 6(19). DOI: 10.1126/sciadv.aaz5373